Check for updates

OPEN ACCESS

EDITED BY Jan-Erik Refle, Université de Lausanne, Switzerland

REVIEWED BY Raluca Popp, University of Kent, United Kingdom María Del Mar Rueda, University of Granada, Spain

*CORRESPONDENCE Alberto Arletti ⊠ alberto.arletti@unive.it

RECEIVED 12 March 2025 ACCEPTED 06 June 2025 PUBLISHED 11 July 2025

CITATION

Arletti A, Tanturri ML and Paccagnella O (2025) Making online polls more accurate: statistical methods explained. *Front. Polit. Sci.* 7:1592589. doi: 10.3389/fpos.2025.1592589

COPYRIGHT

© 2025 Arletti, Tanturri and Paccagnella. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Making online polls more accurate: statistical methods explained

Alberto Arletti*, Maria Letizia Tanturri and Omar Paccagnella

Department of Statistical Sciences, University of Padova, Padova, Italy

Online data has the potential to transform how researchers and companies produce election forecasts. Social media surveys, online panels, and even comments scraped from the internet can offer valuable insights into political preferences. However, such data is often affected by significant selection bias, as online respondents may not be representative of the overall population. At the same time, traditional data collection methods are becoming increasingly cost-prohibitive. In this scenario, scientists need instruments to be able to draw the most accurate estimate possible from samples drawn online. This paper provides an introduction to key statistical methods for mitigating bias and improving inference in such cases, with a focus on electoral polling. Specifically, it presents the main statistical techniques, categorized into weighting, modeling, and other approaches. It also offers practical recommendations for drawing estimates with measures of uncertainty. Designed for both researchers and industry practitioners, this introduction takes a hands-on approach, with code available for implementing the main methods.

KEYWORDS

review, non-probability samples, non-ignorable selection, electoral polling, missingness, MRP

1 Introduction

Random sampling, one of the most powerful tools in scientific research, was first introduced in 1934. The idea is simple. Given a small portion of individuals in a group, it is possible to obtain a reliable estimate for the parameter of interest for the whole population, such as the population mean. A random sample, or probability sample—adjectives which will be used interchangeably in the text—possesses therefore the seemingly magical power of achieving an estimate even with few values of *n*, the sample size, compared to *N*, the population size (Smith, 1976). The key to such a feat of random sampling lies in managing to obtain a sample which is *entirely random* with respect to all aspects that might influence the parameter of interest. To do so, researchers often need to know the probability of each individual in the population to join the survey, a value called inclusion probability. If such value is known and is not zero, the sample can be considered a representative probability sample.

Although straightforward, respecting such a requirement in practice can be a major issue. This is especially true when drawing measurements of complex human phenomena, such as voting behavior. As stated by Kruskal and Mosteller (1979) "the idea will rarely work in a complicated social problem because we always have additional variables that may have important consequences for the outcome" (p. 249).

The delicate complexity of social problems requires random sampling to follow extra steps in order to obtain effective randomness in the sample, and therefore maintain its status as the "dominating" sampling mechanism. For example, when contacting citizens in order to measure their voting intentions for an upcoming election, randomness could be achieved by calling phone numbers at random, given a list of all phone addresses in a given area (method referred to as Random Digit Dialling, or RDD). But what about people who can't answer the phone, for a variety of reasons that *could* be connected with their choice of vote, and therefore generate bias in the final outcome? In other words, "polling of humans is far from the simple random sampling described in many statistics textbooks" (Gelman, 2021, p. 69).

The issue of achieving randomization when human factors are at play is further hampered by another important aspect: declining response rates. The decrease has been abundantly reported (Brick, 2011), with a recent example being the decline from 60% in 2004 to 40% in 2024 in the European Social Survey (European Social Survey, 2024). This decline applies to electoral polls as well (Gelman, 2021). As people are seemingly uninterested in answering field researchers, two consequences appear: Firstly, a non-response bias is introduced in electoral polls (Shirani-Mehr et al., 2018), which is to say that the individuals who do not respond could be systematically different from those who do respond. Secondly, conducting research becomes more expensive, as more and more people need to be contacted to obtain a representative sample (Baker et al., 2013b). Representative surveys can also be more expensive, even without considering response rates. For example, Unangst et al. (2020) reports the cost for a single interview to be 10\$ when conveniently obtained from the internet, with little guarantees of randomness, while the cost climbs to 192\$ for the more selection-safe face-to-face approach. In addition to the prohibitive costs, given the complexity of social sciences and the increasing rates of non-response, one might legitimately question whether a truly random sample is still achievable at all. These two considerations led some researchers to state that there is no such thing as a "random sample" anymore (Bailey, 2023; Beaumont and Haziza, 2022) or, humorously, that "non-random samples are almost everywhere" (Meng, 2018, p. 718). These two consequences have led researchers and polling companies to increasingly turn to alternative methods of sampling. In the following, we introduce non-probability sampling as a pragmatic response to the challenges and limitations of probability-based approaches.

1.1 Non-probability samples

Given the aforementioned problems, researchers might need alternative methods for data collection. To the rescue come nonprobability samples, or non-random samples. Non-probability samples are all samples that come from a vast number of techniques used to obtain data, from snowball sampling (Dusek et al., 2015) to asking people's opinion on social media (Alexander et al., 2020) to scraping web pages (Schirripa Spagnolo et al., 2025), to many others. Such samples are cheaper and more convenient to obtain, and therefore a very popular choice for researchers and practitioners.

In the social sciences, non-probability samples can be advantageous due to their versatility, low cost, and possibility of being employed where other methods often cannot. In particular, speed can be a remarkable quality. For example, the influx of online non-probability data can allow feats such as using Facebook Advertising Platform to nowcast the distribution of migrant groups in the United States, as in Alexander et al. (2020) and Zagheni et al. (2017). In another example, the stream of non-representative Twitter data has been used to provide fast-updated estimates of pre-electoral polls for the US elections (Beauchamp, 2017). Nonprobability samples can also be used to make updated forecasts when more recent census data is unavailable, such as in using Google searches to forecast birth rates (Billari et al., 2016). Continuing, non-random sampling can often be the only viable strategy to examine hard-to-reach populations, as for example using mobile and landline in De Vries et al. (2021), using LinkedIn as in Dusek et al. (2015), or using the social media platforms Vkontakte e Odnoklassniki (Rocheva et al., 2022). Migrants are an especially salient case of such populations, which might not fit in the traditional administrative or random sampling schemes. For example, Zagheni et al. (2014) used localized tweets to draw a nonrandom sample used to infer migration patterns, while Jacobsen and Kühne (2021) used a tracking app for the same aim. Finally, it is clear that using social media, a case of non-probability sampling, offers the advantage of smaller prices and a relatively large pool of individuals to draw from. While most samples obtained online can often be considered non-probabilistic in nature, it is worth noting that some online probability samples exist, as Blom et al. (2016).

1.1.1 Pollings and the shortcoming of non-probability samples

Even though it is clear that non-probability samples such as opt-in online panels or social media data can be a game changer in many scenarios, the significant drawbacks of selection bias, which might result in less accurate results, must be accounted for Callegaro et al. (2014b). Selection bias can be defined as systematic differences between the sampled and target populations, due to the fact that the survey was accessible to a section of the population only, for example, internet users only or Facebook users only. Nonprobability samples are non-representative as they carry selection bias, which leads to a violation of the canon of randomness in some measure.

Because non-probability samples contain this selection, drawn estimates, such as the predicted share of votes for a said party, are not reliable, as in they do not represent the target population of interest, but rather the selected subgroup from data was extracted (e.g., Facebook users who happened to be online at the time of the survey). Therefore, while selection is used in a random sample to select a sample which is random in all its characteristics with respect to the interest statistics, non-random samples are vulnerable to the adverse effect of selection (Kruskal and Mosteller, 1979, p. 246). Some examples of such violations of the pure assumptions of probability sampling are nonresponse, incomplete coverage of the population, and measurement errors (Brick, 2011). The effect is that the "magical" quality of random samples is not applicable anymore, and suddenly the small size n of the sample is unable

to measure correctly the large N of the interest population (Meng, 2022). Therefore, this has led the American Association for Public Opinion Research (AAPOR) in 2010 (Baker et al., 2010) and again in 2013 (Baker et al., 2013a, p. 12) to state that "researchers should avoid non-probability opt-in-panels when a key research objective is to accurately estimate population values... claims of representativeness should be avoided when using these sample sources."

Another issue is that respondents in non-probability surveys, such as those collected via social media or online panels, tend to provide less informative responses compared to more involved methods like face-to-face interviews. For instance, Fricker et al. (2005) and Heen et al. (2014) document "depressed responses" in such settings, evidenced by answer clustering around the middle of the scale, reduced differentiation, and fewer extreme opinions.

Arguably, the field where non-probability samples' shortfalls have generated the strongest shockwave is electoral polling (Evans and Mathur, 2018; Zagheni and Weber, 2015; Shirani-Mehr et al., 2018). As put eloquently in a 2018 review: "Polls have had a number of high-profile misses in recent elections. Political polls have staggered from embarrassment to embarrassment in recent years" (Prosser and Mellon, 2018, p. 757). Famous examples are the 2016 presidential race (Kennedy et al., 2018) [which has been named "a black eye" for polling (Gelman, 2021, p. 67)], the 2016 Brexit referendum (Financial Times, 2016), and the 2023 Turkish general elections (Selcuki, 2023). Generally, the failure of those polls is mainly attributed to the use of non-probability samples (Gelman, 2021), as such samples have been reported as less accurate compared to probability sources (Sohlberg et al., 2017; Sturgis et al., 2018). Nonetheless, the trend does not seem to be stopping for the rise of non-probability samples in electoral polling as well (Callegaro et al., 2014a). A failure in an electoral prediction bears a higher cost for the public image of the discipline. After all, "election polling is arguably the most visible manifestation of statistics in everyday life" (Shirani-Mehr et al., 2018, p. 608). Election polling is almost the most salient because poll-based forecasts are compared to actual election outcomes (Gelman, 2021).

Researchers might end up stuck between a rock and a hard place. Random samples can hardly be completely trustworthy and require heavy costs compared to the cheaper non-probability alternatives (Tam and Clarke, 2015). On the other hand, nonprobability samples carry important challenges for inference. Given these premises, what should researchers do with the abundant quantities of non-random samples available, such as Twitter posts, Google searches, online, and opt-in panels, etc.? It is clear that the need for reliable approaches to draw valuable inferences from non-probability samples is pressing and might bring great benefits to the academic community. After all, "Great advances of the most successful sciences—astronomy, physics, chemistry were and are achieved without probability sampling." (Kish, 1965, pp. 28–29).

From this scenario, the need for statistical methods used to draw valid inferences from non-probability social science data emerges as paramount for the whole scientific community. Statistical methods could aim at reducing or acting as a counterweight to the distortion or bias present in such nonprobability samples. In other words, the estimated value would be closer to the true population value after applying the estimation method, in the form of calibration or correction.

Given the potential of non-probability data sources, such as social media online surveys, for the social sciences and opinion research, such as electoral polling, it is crucial to explore statistical methods that reduce bias and improve accuracy in such datasets. This work aims to assist researchers and practitioners by outlining key statistical techniques for correcting non-probability data, focusing on reducing distortion or bias. It provides an accessible overview of these methods, their assumptions, and practical implementation, serving as a reliable guide for selecting and applying the appropriate approach in their analyses.

2 Data availability scenarios in non-probability sampling

Addressing selection bias in non-probability samples requires appropriate statistical methods, but their applicability depends on the available population information. Researchers may find themselves in different data availability scenarios when working with non-probability samples, which are briefly illustrated here.

In the simplest case, only sample data is available, with no population reference (e.g., hard-to-reach groups like migrants, where census data is lacking). More commonly, researchers also have population totals, as in electoral data, which may be available in marginal (e.g., total voters by sex or region) or cross-tabulated form (e.g., female voters by region). Lastly, some non-probability samples can be paired with a (often smaller-sized) probability sample (Tutz, 2023; Rafei et al., 2022). The present contribution focuses on the second case, where marginal or cross-tabulated totals are available. The first case allows little room for correction, while the third involves distinct challenges and is less common in electoral poll practice.

In the second setting, population information is available as either marginal totals or cross-tabulated census data. This can be represented as a dataset with a target variable *Y*, a set of covariates *X* with *p* parameters, and a *p*-sized vector T(X) containing population totals for each variable in *X*. When complete cross-tabulated census data is available, the researcher has two datasets: (1) A non-representative sample containing *Y* and predictors (also named covariates in the text) *X* (*n* rows). (2) A representative dataset of the full population (*N* rows) with covariates *X*, but without *Y*. These datasets can be concatenated with an indicator variable *S*, where S = 1 for sampled units and S = 0 otherwise (see Figure 1).

An additional important concept is population cells. Any population, such as voters in a country, can be divided into nonoverlapping cells. Each cell represents a unique category in the population, defined by a specific combination of categorical X variables. For example, a cell might be male, 30-45 years old, voter. The total number of cells is given by the product of the levels of available categorical variables. For instance, if gender (2 levels) and employment status (3 levels) are available, the population is divided into $2 \times 3 = 6$ cells. The *X* covariates can also be political affiliation variables in the case of electoral polling, such as party affiliation or the party voted in the previous elections.

Finally, hands-on practice enhances the learning of new methods. To complement the theoretical discussion, this



introduction is accompanied by a sample dataset and code implementations for most methods presented. This allows readers to grasp both the technical details and practical application. The code and data are available on GitHub: nonign_sel_companion.

3 Weighting

Weighting, or calibration weighting, first introduced by Deville and Särndal (1992), is considered one of the most important methods for correcting a non-representative sample (Valliant, 2020). In weighting, the individual observations are up or down weighted so their distribution is adapted to be more similar to the distribution of a representative sample or of the census. In their most basic idea, if the sample has way more males than females compared to the known national totals, then male observations can be down-weighted. This class of methods can also be referred to as "pseudo-weighting" or "quasi-randomization" (Valliant, 2020). This is due to the fact that in random sampling, observations in the sample are weighted by the inverse of their inclusion probability, which is known (see Horvitz and Thompson, 1952). In the case of non-random sampling, the inclusion probabilities are not known and are to be estimated. Therefore, weighting is used in trying to approximate sampling weights in a manner that resembles what is done in probability sampling. In the case of unknown inclusion probabilities, or non-random samples, weighting can be obtained with one, or a combination of raking, propensity scoring, and matching.

3.1 Raking (iterative proportional fitting)

Iterative Proportional Fitting, or Raking, Deming and Stephan (1940) is a weighting method which is used to weight a dataframe

so that the X variables' marginals match the corresponding population marginals. This is done in the case of multiple marginal distributions, for example, gender and region. The term iterative is used to refer to the process that is used to obtain the weighting, which can be described in simple words as adjusting the weights iteratively, making them more similar to the marginals at each iteration until convergence (Stephan, 1942).

The goal of raking is to assign weights $w_1 \dots w_j \dots w_n$ to each row in the sample so that the weighted sums match known population totals from the census. For the covariate p, this can be expressed as:

$$\sum_{j=1}^{n} w_j x_{j,p} = T(X_p).$$
(1)

Here, $T(X_p)$ is the population total for the *p*-th covariate, and $x_{j,p}$ represents the value of the *p*-th covariate for row *j*. The estimated population mean $(\hat{\mu}(Y))$ is then obtained as:

$$\widehat{\mu}(Y) = \frac{1}{n} \sum_{j=1}^{n} y_j w_j.$$
⁽²⁾

This formula allows the estimation of the population total for the target variable, such as the share of votes. If one would like to obtain measures of uncertainty around such an estimate, a common practice is to use a bootstrap or similar resampling approaches (Kolenikov, 2010). Alternatively, a direct expression to obtain raking variance is provided in Deville and Särndal (1992).

Raking is a very simple weighting method that only requires the marginal distributions to be employed, and is especially useful in the case where only marginal totals are available (see Section 2), or in cases where the number of observations in each cell is small. Nonetheless, it can suffer from a series of limitations. To begin with, the raking to the marginals does not take into account possible higher-level interactions between the raking variables. This can be an issue, making the weighing less accurate compared to the real population distribution. A proposed solution to this problem is multilevel calibration weighting, an approach by Ben-Michael et al. (2021). While raking matches the marginal distributions of the raked variables only, it might not be able to balance higher-order interactions. Multilevel calibration weighting aims at solving that, behaving similarly to raking but adding some approximate balance for interaction, prioritizing lower-order interactions. In addition, if the raking variables do not fully account for the inclusion probability, the method becomes inconsistent. Finally, it should be noted that the weights produced by raking can have very high (or very low) values, making the practice unreliable. One possible solution is "trimming" the weights, or constraining the weights to be in a certain range of values. Such a solution is also implemented directly in the R command for raking anesrake (Pasek, 2018).

3.2 Propensity score adjustment

Propensity Score Adjustment is a class of adjustment methods that relies on the estimation of the probability of inclusion in the non-probability sample. The main method, discussed here, is often referred as Propensity Score-based Inverse Probability Weighting (PS-IPW) (Zou et al., 2016).

PS-IPW works through the use of a second representative sample, with common covariates to the non-probability sample, where the target *Y* variable is missing (Schonlau and Couper, 2017; McPhee et al., 2022). Such a sample can be generated knowing the census cross-tabulated totals, if those are available. To do so, it would be sufficient to generate a dataframe where each column corresponds to a census cross-tabulated variable, and the number of rows that belong in each cell corresponds to, or is proportional to, the known population total. The two datasets are temporarily binned into a single frame, as described in Section 2 and Figure 1. Then, the method builds a weighted logistic regression model to estimate the probability of an observation being in a nonprobability sample. Here, the regression weights correspond to the known inclusion probabilities in the reference sample, while nonsampled observations receive a weight of 1. Inclusion probabilities in the reference sample correspond to the known probability of individual *j* of being included in the sample, which generally accompany a representative sample. If the reference sample has been generated from the cross-tabulated census values, then the inclusion probability of a row *j* belonging to cell *c* is just the inverse of the numerosity of that cell. The regression can be described as:

$$logit(P(S = 1|X)) = \beta_0 + \beta^T X.$$
(3)

The predicted values of the weighted regression, which can be set as $\hat{\pi}_j = \text{logit}(P(S_j = 1|x_j))$, are then inverted and used to estimate the population mean.

$$\widehat{\mu}_{\text{PS-IPW}}(Y) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{\widehat{\pi}_j} y_j.$$
 (4)

This last formula is the same as the famous Horvitz-Thompson estimator (Horvitz and Thompson, 1952), with the difference that the weights are not known from the sample design, but are estimated from the data. It is also similar to Equation 2, with the difference that the $\frac{1}{\hat{\pi}_j}$ values are estimated differently from w_j . Such a probability estimated from the data is called the propensity score. What this achieves is an estimation of the inclusion probabilities, which are unknown, from the observed data. The propensity score represents the conditional probability of being included in the survey given an individual's covariate profile.

For a measure of uncertainty of this estimate, variance estimates can be obtained through a Taylor linearization approximation (Valliant et al., 2013, p. 426) or through a Jackknife approximation (Valliant, 2020, p. 8).

One topic of discussion regards which model to choose in order to obtain propensity scores. While logistic regression is a very popular method, some authors argue that it is insufficient in cases such as where the propensity score shows a non-linear function. In this regard, Lee et al. (2010) compares the performance of different methods to obtain propensity scores. They compare both logistic regression with Classification and Regression Trees (CART) models, and find that logistic regression's performance can deteriorate in case of non-additivity and non-linearity. Therefore, choosing a flexible or non-parametric approach to model propensity scores can be advantageous. For example, Rafei et al. (2020) uses Bayesian Additive Regression Trees (BART) to model inclusion probabilities (see also Elliott et al., 2010 for Bayesian modeling of this kind). Furthermore, attention should be dedicated to choosing the appropriate variables to correctly model inclusion probabilities. To this task, variable selection methods such as in Ferri-García and Rueda (2022) can be employed.

Once an appropriate model has been selected to capture the selection mechanism and there are no empty cells in the data, the PS-IPW can be used to build reliable estimators. A first assumption of this approach is that every unit in the population has a non-zero propensity score. A second important assumption is that the covariates X should include all relevant confounders (Lee and Valliant, 2009). The main danger in using this method emerges when the selected X variables do not fully account for the sample selection mechanism, or in other words, there is significant selection bias that cannot be controlled by the available covariates. In that case, adjusting for the propensity score will not produce unbiased estimates of the treatment effect. A further requirement of PS-IPW is called "common support" and requires that the distribution of the covariates in the reference sample is similar to the distribution in the sample to be adjusted. For example, there should not be population cells completely absent from the nonprobability sample (Valliant, 2020). Pseudo-inclusion probabilities are typically estimated using weighted logistic regression (Lee and Valliant, 2009).

4 Modeling

Another popular approach to adjust non-probability surveys and reduce selection bias is modeling. In this case, the non-random sample is employed to train a model used to predict the dependent variable for each cell of the missing rows, corresponding to the population. This approach is also called superpopulation model estimation (Valliant, 2020), model-based predictive inference (Buelens et al., 2018), or model-based estimation (Wu, 2022). In modeling, the y_i values of the non-sampled units are predicted with a variety of methods trained on the sampled units. In this way, the value for the total population is considered the union of both the sampled and non-sampled units. That is, the non-sampled units correspond to all individuals who are in the target population, but not in the sample.

4.1 Post-stratification

Superpopulation methods are therefore comprised of two steps, a first modeling step, where the model is estimated from the observed data, and a post-stratification step, where the value is predicted for each cell of the population. The sum of all predicted values for all cells gives the estimated value for the entire population. After modeling, the post-stratification step allows for balancing for sample discrepancy.

An estimate of the population mean for a given cell of the population, y_c , where the subscript *c* indicates the cell, can be obtained by first estimating a model between *Y* and *X* in the sample,



for example, a linear regression. This can be described as:

$$logit(P(Y = 1|X)) = \beta_0 + \beta^1 X.$$
(5)

Then, the cell total can be obtained using the following formula:

$$\widehat{\mu}(Y_c) = \frac{N_c}{N} (\beta_0 + \beta^T X_c) \tag{6}$$

where N_c is the known size of the population cell c, X_c indicates the matrix of covariates data for the cell c, and $\hat{\beta}$ are the estimated regression coefficients. The estimate for the whole population will be the sum of all cell totals, so that $\hat{\mu}(Y) = \sum_c \hat{\mu}(Y_c)$.

While the post-stratification adjustment step remains the same across applications, what can be changed is the model used for prediction. A simple linear model can be substituted with more complicated or non-linear models. In this regard, Ferri-García et al. (2021) and Castro-Martín et al. (2020) examine the use of machine-learning models as prediction models, such as neural networks and decision trees. However, when predictors are demographic categorical variables, a hierarchical model is most effective, and such adjustment is referred to as Multilevel Regression and Post-stratification (MRP). MRP is not of recent development, with Gelman (1997) being the original proposer of the method. Nonetheless, MPR is one superpopulation method that is frequently used with non-representative surveys (McPhee et al., 2022; Si, 2020). In MRP, a multilevel regression model is used to estimate the outcome variable using a larger number of auxiliary variables and their interactions than is possible with standard weighting methods. The particularity of MRP is that it performs a cell-based (sub-group) estimation, and the hierarchical component (with Bayesian prior in its original specification, see Li and Si, 2022) regularizes the model and allows for borrowing of information.

MRP is a key method in the field, and it provides several advantages over post-stratification with a simple linear regression. To best understand the mechanics of MPR, it can be useful to examine the following formula for estimating the population mean using MPR (Si, 2020, p. 5):

$$\widehat{\mu}_{\mathrm{MRP}}(Y) = \sum_{c} \frac{N_{c}}{N} \frac{\widehat{\mu}(Y_{c}) + \delta_{c}\widehat{\mu}(Y)}{1 + \delta_{c}}, \quad \text{where} \quad \delta_{c} = \frac{\sigma_{c}^{2}}{n_{c}\sigma_{Y}^{2}}.$$
 (7)

Here, as in Equation 6 the subscript *c* indicates a post-stratification cell, $\hat{\mu}(Y_c)$ is the model estimate for cell *c*, N_c is the size of cell *c* in the population, $\hat{\mu}(Y)$ is the estimated population mean, σ_c^2 is the variance of the outcome variable for cell *c*, n_c is the sample size for cell *c* and σ_Y^2 is the outcome between-cell variance. Between-cell variance is a measure of how much the mean of *Y* differs from one cell to another, reflecting systematic differences between groups defined by the stratifying variables (e.g., age, gender, region). Therefore, what this formula tells us is that the less information we have on cell *c*, both in terms of sample size and variety, the more we are going to "borrow" from the other cells. This method is especially effective in non-probability online panel samples or social media samples, where it's quite often the case to have cells with very few observations.

For uncertainty measures on the population estimates of both post-stratification and MRP usually a Bayesian approach with posterior draws is usually preferred (Lopez-Martin et al., 2022).

For illustration purposes, an example of an electoral poll adjustment using Bayesian MRP is presented in Figure 2. In the plot, the blue dotted and dashed line represents the unadjusted sample mean, while the red dotted line represents the true population value for the share of votes of the center-left coalition in the 2022 Italian elections. The black dashed line represents instead the adjusted population estimate using MRP. In each subplot, the marginal probability of each subpopulation cell to vote for that party is plotted, together with credibility interval bands.

It has been noted that post-stratification is useful in reducing selection bias and correcting imbalances in the sample composition. One advantage of such estimators is their ability to reduce bias (Kim et al., 2021). In this regard, the method has shown to be capable of impressive bias-correcting performances in election forecasting, for example in Wang et al. (2015). Nonetheless, when drawing inference with such method, some factors come into play to determine its performance. The first factor is the need for high-quality predictive post-stratification variables, or, in other words, variables with a strong relationship with the outcome variable. Authors have reported how poorly predictive auxiliary information might have an important effect on the final outcome (Si, 2020), and that variables chosen for post-stratification are more relevant than the model used for estimation (Prosser and Mellon, 2018). For example, Buttice and Highton (2013) examines the correlates of MRP performance in various scenarios. The authors examine how MRP accuracy of estimates of election results varies as the strength of the relationship between voting opinion and state-level covariates increases. They observe that as the strength of the relationship between opinion and the state-level covariates increases, then also MRP estimates get closer to the true values. This is not seen with the same strength for the individual-level covariates.

The requirement for high-quality post-stratification variables can be challenging when the census is limited. The requirement to have cross-tabulated population tables can be daunting, especially as the number of covariates increases. Therefore, it is often the case that variables useful for adjustment are not included in the census, such as party identification or previous vote (Gelman, 2021). Usually, due to non-availability in census, post-survey adjustments are limited to basic demographics such as age, gender, race, and education from large-scale government surveys (Chen et al., 2019). Moreover, for the case of electoral polling, these problems can be exacerbated for practitioners working outside of the United States. For example, pollsters in the United States can access party registration information, which is generally unavailable in other countries (Prosser and Mellon, 2018). Therefore, MRP has generally been applied so far in election forecasts for a few countries (Leemann and Wasserfallen, 2017). As a possible solution, Kastellec et al. (2015) suggests expanding the post-stratification table by incorporating a survey that includes one or more non-census variables, which can aid in adjusting for discrepancies between the sample and the target population. Such practice can be referred to as "embedded MRP" or e-MRP (Li and Si, 2024; Ornstein, 2023).

5 Other methods

5.1 Statistical matching

Statistical Matching, also known as Sample Matching or Mass Imputation, is a technique that can be applied both before the sample is selected (Cornesse et al., 2020; Bethlehem, 2016) or after the non-probability sample is already obtained (Mercer et al., 2018). The approach for the second case, the one of interest for the purpose of the present work, is attributed to Rivers (2007). Similarly to Propensity Score Adjustment, it requires a probability sample where the target variable does not need to be measured, but where there are matching covariates. The reference sample is treated as a target, where each row of the target is paired with the closest observation in the non-probability sample. The "matching" observation is chosen to be an observation that has the strongest similarity in the covariates. A Euclidean distance metric can be used (Cornesse et al., 2020), as well as any sort of similarity matrix, such as one obtained from a Random Forest (Mercer et al., 2018). Alternatively, a nearest neighbor approach can be useful, especially in the cases of continuous variables or categorical variables with many ordinal levels (Chen and Shao, 2000). The closest match is chosen for each row of the reference sample, and any remaining observation that has not been paired is discarded. Sequentially, each observation in the target dataframe is matched one at a time, and the most similar case is chosen among the cases which has not been matched previously. Then, the statistics of interest are obtained using the target variable y of the matched cases. In other words, each row of the target reference sample is substituted with the most similar observation in the non-probability case.

The main limitation of matching is that, in order to obtain a meaningful matching, a sufficiently large set of variables should be available in the required probability sampling. Most often, these variables should be different than the common demographic variables and might not be present in the available census. Otherwise, other forms of adjustment would be more straightforward. For the case of electoral polling, obtaining a reference sample with such characteristics can be challenging.

5.2 Inverse sampling

Inverse Sampling is presented for the estimation of nonprobability big data samples in Kim and Wang (2019). The idea of inverse sampling is to leverage the large n of the non-probability sample to make a sub-selection. The first-phase sample consists of big data, named A, which is affected by selection bias. The secondphase sample, named A_2 , is a subset of the first-phase sample, designed to adjust for this selection bias. To extract the subsample, inclusion probabilities proportional to the importance weights are used for selection. External information from a reference sample or from the census is used to correct for selection bias in the second step.

5.3 Doubly-robust estimation

Doubly-Robust estimation or Doubly-Robust Poststratification (DRP) is substantially a combination between weighting, seen in Section 3, and modeling, seen in Section 4. The fundamental idea is to combine the two components, a propensity score component and modeling with a poststratification component. When estimating a propensity score model, the specified model might be incorrect; for example, it might ignore interactions that influence the selection mechanism. The same might be for the modeling approach, where the chosen

10.3389/fpos.2025.1592589

model might not be the best fit to describe the relationship between the target variable (Y) and the available covariates (X) (Tan, 2007). In DRP, the final estimate will be correct as the sample size increases even if one of the two models, either the modeling or the propensity score model, is incorrect or misspecified (Theorem 2; Chen et al., 2020). This guarantees further protection against bias. Similarly to DR-IPW, imagine a second reference sample where Yis missing. We call the non-probability sample A and the reference probability sample B. To obtain DRP, two models are fitted,

- 1. A propensity score model on the probability of *j*-th being included in *A*, using, for example, a weighted logistic regression as in Equation 4. The predicted propensity score is again $\hat{\pi}_j$ for row *j*.
- 2. A model of the relationship between the target *Y* and the covariates *X*, using *A* data only, as in Equation 5. The predicted value of *y* for row *j* by this model is indicated as \hat{y}_j .

For the case of a linear model, the final DRP population estimate is obtained by:

$$\hat{\mu}_{\mathrm{DR}}(Y) = \frac{1}{\sum_{i \in A} 1/\hat{\pi}_i} \sum_{i \in A} \frac{1}{\hat{\pi}_i} (y_i - \hat{y}_j) + \frac{1}{N} \sum_c N_c(\beta_0 + \beta^T X_c).$$
(8)

Unpacking this expression, $\sum_{i \in A}$ indicates to sum across all rows in the *A* dataframe, while \sum_c for each population cell. The first term in Equation 8 sums the difference between the predicted and the measured values of *Y* for the non-probability sample, weighted by the inverse of the probability score obtained with the propensity score model. The second term is a post-stratification, as in Equation 6.

An estimate of variance of the DRP is present in Chen et al. (2020), but bootstrap resampling can also be used (e.g., see Beresewicz and Szymkowiak, 2024). Point estimation of DRP in R can be easily carried out with the nonprobsvy and Non-ProbEst packages (Chrostowski and Beraesewicz, 2024; Rueda et al., 2020), for example, which also provide estimates of the uncertainty of the predicted population mean. For a list of R packages to this aim, the reader is directed to Cobo et al. (2024). The method has a set of strong qualities on paper, but real-world application might vary widely (Si, J., personal communication, 11/2023). This might be due to the fact that the two model components' effects might interact with one another, creating either a more unpredictable behavior (Meng, X. L., personal communication, 11/2023). In conclusion, DRP offers notable advantages theoretically, but has not yet replaced other methods in practical applications automatically.

6 Limits of the presented approaches

All the models presented in the previous sections assume that the selection mechanism is entirely explained by the X covariates alone. If the selection mechanism is not entirely explained by X, then the estimated model might not provide accurate estimates of the population of interest. Importantly, there is abundant evidence that non-probability samples might suffer from nonignorable selection, or in other words, that S is not only influenced by X alone but by the target variable Y as well. In political polling, this might be the case due to a variety of reasons. For example, respondents in non-probability panels being generally more politically engaged than the general population (Prosser and Mellon, 2018), respondents who vote for a candidate who is doing well might be more likely to answer a survey (Gelman et al., 2016), ads being used to target responders failing to be neutral or to attract voters of a specific political affiliation (Matz et al., 2017; Zarouali et al., 2022; Schneider and Harknett, 2022; Kühne and Zindel, 2020), online responders having different personality characteristics compared to the global population (Valentino et al., 2020; Brüggen and Dholakia, 2010), or online samples having no respondents in certain cells of the population (Bartoli et al., 2019). Despite the many mechanisms that can lead to non-ignorability in online samples, available methods to address this problem are not widely diffused. Examples of useful approaches in this regard are Burakauskaitė and Čiginas (2023) or Marella (2023). One field where methods might be applied to this case is missing data theory, where mechanism missingness can be considered the same driving selection, simply inverted. In this case, some reweighing methods have been proposed to adjust for non-ignorable missingness (for example, see Matei, 2018), as well as models which use assumptions on the selection mechanism to adjust for selection bias (see West and Andridge, 2023 and Andridge, 2024).

All in all, while the methods presented here might prove to be capable in reducing selection bias from samples collected online, researchers should be conscious that some selection mechanisms cannot be completely undone without stronger assumptions or knowledge of the sampling mechanism.

7 Conclusions

This paper reviewed the main methods used for adjusting a non-probability sample, such as an online sample, with a focus on electoral polling. While each method has been described in general terms, the choice of which one to use in each situation can depend on the specific setting, data availability, and research goal. One useful resource in this regard is Cornesse et al. (2020), which also had a setting centered on non-probability samples used to estimate election polls. The authors compare probability samples with corrected or weighted non-probability samples. They compare some approaches listed in the previous sections: (a) Calibration weighting using post-stratification or raking; (b) Sample matching; (c) Propensity score weighting; (d) Pseudo-design based estimation such as propensity score weighting; They find that weighting can reduce the bias in some cases, but in general the authors arrived to the conclusion that weighting does not suffice in completely eliminating bias in non-probability based surveys.

One general rule that can be applied to all methods is that as long as strong predictive variables are available, in weighting or in modeling alike, most of the selection mechanisms can be accounted for. As X decreases in predictive power, things get more complicated: selection might be unaccounted for, and researchers have fewer tools at their disposal in obtaining an estimate. Concluding, we go back to an important concept expressed in the introduction: the large n (typical of non-probability samples), is, alone, unable to provide unbiased estimation. Nonetheless, a rich X, or a wide dataset of covariates, might instead be a more fruitful pathway toward robust estimation. In this sense, to work

TABLE 1 Summary of adjustment methods pros and cons.

Method	Pros	Cons	Use when
Raking	Can be used when population marginals only are available.	Does not capture higher-order interaction.	No cross-tabulated data is available.
PS-IPW	Flexible, large availability of models to estimate weights.	Propensity model must be correctly specified, requires cross-tabulated population table.	Modeling the inclusion <i>S</i> variable is priority.
Model-based inference	Large availability of models.	Requires cross-tabulated data, model must be correctly specified.	Modeling the <i>Y</i> <i>X</i> relationship is priority.
DRP	Only the inclusion model or the outcome model needs to be correctly specified.	Results can vary widely from application to application.	When data and variables are rich and of good quality, but there is uncertainty in which is the correct model for <i>S</i> or for $Y X$.

well, non-probability online samples should not just be big, but rich as well. Techniques that might be the most promising in this sense are therefore the ones which allow for an expansion of prediction variables, such as in Li and Si (2024); Kuriwaki et al. (2024), and methods that allow the researcher to add previous knowledge on the possible selection mechanism, such as in Little et al. (2020). In general, estimation with non-probability samples in electoral polling should proceed carefully, depending on the selection mechanism.

For this reason, it is difficult to give general recommendations on when to use this or that other method. Mostly, the literature points out to the fact that variables, rather than the chosen adjustment method, have the lion's share in making the adjustment effective (Little and Vartivarian, 2005; Elliott and Valliant, 2017; Gelman, 2007; Rafei et al., 2020; Mercer et al., 2018; Prosser and Mellon, 2018). Nonetheless, as a general rule, a few directions can be indicated to guide a researcher. If only the population marginal totals are available, then raking can be a robust adjustment option. When cross-tabulated population totals are available, both propensity score based methods and predictive modeling methods are valid, where the first concentrates more on modeling the S selection mechanism while the latter the Y|X mechanism, so the choice between the two should be guided by considering weather the data and variables might be more informative on one or the other mechanism. Finally, DRP is also a useful approach, especially in the case where variables are strongly predictive of both the S and Y|X mechanisms, but the researcher is not certain of the shape of the relationship. With proper caution and consideration of the factors discussed earlier, readers may refer to Table 1 for a summary of the key use cases for each statistical method.

While non-probability samples pose significant challenges due to selection bias, they also offer valuable opportunities when handled with the right statistical methods. This paper has provided both an intuitive and technical overview of key approaches to adjust for bias and improve inference. Although no method can fully replace probability sampling, the techniques discussed here can enhance the reliability of estimates derived from nonrepresentative data. By increasing awareness of both the risks and potential of these samples, this work aims to support researchers in making informed methodological choices when working with online and other non-probability datasets.

Author contributions

AA: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. MLT: Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing. OP: Funding acquisition, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was partfunded by the PON "Research and Innovation" 2014–2020 Actions IV.4 "PhDs and research contracts on innovation issues" and Action IV.5 "PhDs on Green issues." Ministerial Decree 1061/2021 as a PhD studentship to Alberto Arletti. Open Access funding provided by Università degli Studi di Padova | University of Padua, Open Science Committee The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative Al statement

The author(s) declare that Gen AI was used in the creation of this manuscript. Open AI GPT-4 is used to edit the manuscript, check grammar and spelling mistakes and latex formatting Open AI GPT-4.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alexander, M., Polimis, K., and Zagheni, E. (2020). Combining social media and survey data to nowcast migrant stocks in the United States. *Popul. Res. Policy Rev.*1–28. doi: 10.1007/s11113-020-09599-3

Andridge, R. R. (2024). Using proxy pattern-mixture models to explain bias in estimates of covid-19 vaccine uptake from two large surveys. J. R. Stat. Soc. Series A 187, 831–843. doi: 10.1093/jrsssa/qnae005

Bailey, M. A. (2023). A new paradigm for polling. *Harvard Data Sci. Rev.* 5:9898. doi: 10.1162/99608f92.9898eede

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., et al. (2010). Research synthesis: AAPOR report on online panels. *Public Opin. Q.* 74, 711–781. doi: 10.1093/poq/nfq048

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., et al. (2013a). *Report of the AAPOR task force on non-probability sampling*. Technical report, American Association for Public Opinion Research (AAPOR).

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., et al. (2013b). Summary report of the AAPOR task force on non-probability sampling. *J. Surv. Stat. Methodol.* 1, 90–143. doi: 10.1093/jssam/smt008

Bartoli, B., Fornea, M., and Respi, C. (2019). "Selection bias and representation of research samples: the effectiveness of mixing mode and sampling frames," in *Poster presented at the 2019 GOR Conference*.

Beauchamp, N. (2017). Predicting and interpolating state-level polls using Twitter textual data. *Am. J. Pol. Sci.* 61, 490–503. doi: 10.1111/ajps.12274

Beaumont, J.-F., and Haziza, D. (2022). Statistical inference from finite population samples: a critical review of frequentist and Bayesian approaches. *Canad. J. Stat.* 50, 1186–1212. doi: 10.1002/cjs.11717

Ben-Michael, E., Feller, A., and Hartman, E. (2021). Multilevel calibration weighting for survey data. *Polit. Anal.* 32, 65–83. doi: 10.1017/pan.2023.9

Beresewicz, M., and Szymkowiak, M. (2024). Inference for non-probability samples using the calibration approach for quantiles. *arXiv preprint arXiv:2403.09726*.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching? Soc. Sci. Comput. Rev. 34, 59–77. doi: 10.1177/0894439315573926

Billari, F., D'Amuri, F., and Marcucci, J. (2016). "Forecasting births using google," in *Carma 2016: 1st International Conference on Advanced Research Methods in Analytics* (Editorial Universitat Politécnica de Valéncia), 119.

Blom, A. G., Bosnjak, M., Cornilleau, A., Cousteaux, A.-S., Das, M., Douhou, S., et al. (2016). A comparison of four probability-based online and mixed-mode panels in Europe. *Soc. Sci. Comput. Rev.* 34, 8–25. doi: 10.1177/0894439315574825

Brick, J. M. (2011). The future of survey sampling. *Public Opin. Q.* 75, 872–888. doi: 10.1093/poq/nfr045

Brüggen, E., and Dholakia, U. M. (2010). Determinants of participation and response effort in web panel surveys. J. Inter. Market. 24, 239–250. doi: 10.1016/j.intmar.2010.04.004

Buelens, B., Burger, J., and van den Brakel, J. A. (2018). Comparing inference methods for non-probability samples. *Int. Statist. Rev.* 86, 322-343. doi: 10.1111/insr.12253

Burakauskaitė, I. and Čiginas, A. (2023). On using a non-probability sample for the estimation of population parameters. *Lietuvos Matematikos Rinkinys* 64, 1–11. doi: 10.15388/LMR.2003.33587

Buttice, M. K., and Highton, B. (2013). How does multilevel regression and poststratification perform with conventional national surveys? *Polit. Anal.* 21. doi: 10.1093/pan/mpt017

Callegaro, M., Baker, R. P., Bethlehem, J., Göritz, A. S., Krosnick, J. A., and Lavrakas, P. J. (2014a). *Online Panel Research: A Data Quality Perspective*. New York: John Wiley Sons. doi: 10.1002/9781118763520

Callegaro, M., Villar, A., Yeager, D., and Krosnick, J. A. (2014b). "A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples," in *Online Panel Research: a Data Quality Perspective*, 23–53. doi: 10.1002/9781118763520.ch2

Castro-Martín, L., Rueda, M., d. M., and Ferri-García, R. (2020). Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques. *Mathematics* 8:879. doi: 10.3390/math8060879

Chen, J., and Shao, J. (2000). Nearest neighbor imputation for survey data. J. Off. Stat. 16:113. doi: 10.1093/jssam/smae048

Chen, J. K. T., Valliant, R. L., and Elliott, M. R. (2019). Calibrating non-probability surveys to estimated control totals using lasso, with an application to political polling. J. R. Stat. Soc. Series C 68, 657–681. doi: 10.1111/rssc.12327

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. J. Am. Stat. Assoc. 115, 2011–2021. doi: 10.1080/01621459.2019.1677241

Chrostowski, L., and Beraesewicz, M. (2024). nonprobsvy: Inference Based on Non-Probability Samples. R package version 0.1.0. doi: 10.32614/CRAN.package.nonprobsvy

Cobo, B., Ferri-García, R., Rueda-Sánchez, J. L., and del Mar Rueda, M. (2024). Software review for inference with non-probability surveys. *Surv. Statist.* 90, 40–47.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., et al. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *J. Surv. Statist. Methodol.* 8, 4–36. doi: 10.1093/jssam/smz041

De Vries, L., Fischer, M., Kroh, M., Kühne, S., and Richter, D. (2021). "Design, nonresponse, and weighting in the 2019 sample q (queer) of the socio-economic panel," in *SOEP Survey Papers* 940.

Deming, W. E., and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Mathem. Statist.* 11, 427–444. doi: 10.1214/aoms/1177731829

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. J. Am. Stat. Assoc. 87, 376–382. doi: 10.1080/01621459.1992.10475217

Dusek, G., Yurova, Y., and Ruppel, C. P. (2015). Using social media and targeted snowball sampling to survey a hard-to-reach population: a case study. *Int. J. Doctoral Stud.* 10:279. doi: 10.28945/2296

Elliott, M. R., Resler, A., Flannagan, C. A., and Rupp, J. D. (2010). Appropriate analysis of Ciren data: using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes. *Accid. Anal. Prev.* 42, 530–539. doi: 10.1016/j.aap.2009.09.019

Elliott, M. R., and Valliant, R. (2017). Inference for nonprobability samples. *Statist. Sci.* 32, 249–264. doi: 10.1214/16-STS598

European Social Survey (2024). Modes of data collection: the ESS move to self-completion data collection. Available online at: https://europeansocialsurvey.org/ methodology/methodological-research/modes-data-collection (Accessed October 16, 2024).

Evans, J. R., and Mathur, A. (2018). The value of online surveys: a look back and a look ahead. *Internet Res.* 28, 854–887. doi: 10.1108/IntR-03-2018-0089

Ferri-García, R., Castro-Martín, L., and del Mar Rueda, M. (2021). Evaluating machine learning methods for estimation in online surveys with superpopulation modeling. *Math. Comput. Simul.* 186, 19–28. doi: 10.1016/j.matcom.2020.03.005

Ferri-García, R., and Rueda, M. d. M. (2022). Variable selection in propensity score adjustment to mitigate selection bias in online surveys. *Statistical Papers* 63, 1829–1881. doi: 10.1007/s00362-022-01296-x

Financial Times (2016). *Brexit poll tracker*. Available online at: https://ig.ft.com/ sites/brexit-polling/ (Accessed October 11, 2024).

Fricker, S., Galesic, M., Tourangeau, R., and Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opin. Q.* 69, 370–392. doi: 10.1093/poq/nfi027

Gelman, A. (1997). Poststratification into many categories using hierarchical logistic regression. *Surv. Methodol.* 23:127.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. Statist. Sci. 22, 153-164. doi: 10.1214/088342306000000691

Gelman, A. (2021). Failure and success in political polling and election forecasting. *Statist. Public Policy* 8, 67–72. doi: 10.1080/2330443X.2021.1971126

Gelman, A., Goel, S., Rivers, D., and Rothschild, D. (2016). The mythical swing voter. *Quart. J. Polit. Sci.* 11, 103–130. doi: 10.1561/100.00015031

Heen, M., Lieberman, J. D., and Miethe, T. D. (2014). A comparison of different online sampling approaches for generating national samples. *Center Crime Just. Policy* 1, 1–8.

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. 47, 663–685. doi: 10.1080/01621459.1952.10483446

Jacobsen, J., and Kühne, S. (2021). Using a mobile app when surveying highly mobile populations: panel attrition, consent, and interviewer effects in a survey of refugees. *Soc. Sci. Comput. Rev.* 39, 721–743. doi: 10.1177/0894439320985250

Kastellec, J. P., Lax, J. R., Malecki, M., and Phillips, J. H. (2015). Polarizing the electoral connection: partisan representation in supreme court confirmation politics. *J. Polit.* 77, 787–804. doi: 10.1086/681261

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J. D., Durand, C., Franklin, C., et al. (2018). An evaluation of the 2016 election polls in the united states. *Public Opin. Q.* 82, 1–33. doi: 10.1093/poq/nfx047

Kim, J. K., Park, S., Chen, Y., and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. J. R. Stat. Soc. Series A 184, 941–963. doi: 10.1111/rssa.12696

Kim, J. K., and Wang, Z. (2019). Sampling techniques for big data analysis. *Int. Stat. Rev.* 87, S177–S191. doi: 10.1111/insr.12290

Kish, L. (1965). Survey Sampling. New York, London: John Wiley Sons.

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. *Stata J.* 10, 165–199. doi: 10.1177/1536867X1001000201

Kruskal, W., and Mosteller, F. (1979). Representative sampling, III: the current statistical literature. Int. Stat. Rev. 47, 245–265. doi: 10.2307/1402647

Kühne, S., and Zindel, Z. (2020). "Using Facebook and Instagram to recruit web survey participants: a step-by-step guide and application," in *Survey Methods: Insights from the Field (SMIF)*.

Kuriwaki, S., Ansolabehere, S., Dagonel, A., and Yamauchi, S. (2024). The geography of racially polarized voting: calibrating surveys at the district level. *Am. Polit. Sci. Rev.* 118, 922–939. doi: 10.1017/S0003055423000436

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Stat. Med.* 29, 337–346. doi: 10.1002/sim.3782

Lee, S., and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociol. Methods Res.* 37, 319–343. doi: 10.1177/0049124108329643

Leemann, L., and Wasserfallen, F. (2017). Extending the use and prediction precision of subnational public opinion estimation. *Am. J. Pol. Sci.* 61, 1003–1022. doi: 10.1111/ajps.12319

Li, K., and Si, Y. (2022). Embedded multilevel regression and poststratification: model-based inference with incomplete auxiliary information. *arXiv preprint arXiv:2205.02775*.

Li, K., and Si, Y. (2024). Embedded multilevel regression and poststratification: model-based inference with incomplete auxiliary information. *Stat. Med.* 43, 256–278. doi: 10.1002/sim.9956

Little, R. J., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Surv. Methodol.* 31:161.

Little, R. J., West, B. T., Boonstra, P. S., and Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *J. Surv. Stat. Methodol.* 8, 932–964. doi: 10.1093/jssam/smz023

Lopez-Martin, J., Phillips, J. H., and Gelman, A. (2022). *Multilevel regression and poststratification case studies*. Available online at: https://juanlopezmartin.github.io, 902,903 (Accessed May 20, 2025).

Marella, D. (2023). Adjusting for selection bias in nonprobability samples by empirical likelihood approach. J. Off. Stat. 39, 151–172. doi: 10.2478/jos-2023-0008

Matei, A. (2018). On some reweighting schemes for nonignorable unit nonresponse. *Surv. Statist.* 77, 21–33.

Matz, S. C., Kosinski, M., Nave, G., and Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proc. Nat. Acad. Sci.* 114, 12714–12719. doi: 10.1073/pnas.1710966114

McPhee, C., Barlas, F., Brigham, N., Darling, J., Dutwin, D., Jackson, C., et al. (2022). Data quality metrics for online samples: Considerations for study design and analysis. AAPOR Task Force Report.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election. *Ann. Appl. Stat.* 12, 685–726. doi: 10.1214/18-AOAS1161SF

Meng, X.-L. (2022). Comments on "statistical inference with non-probability survey samples"-miniaturizing data defect correlation: a versatile strategy for handling non-probability samples. *Surv. Methodol.* 48, 339–360.

Mercer, A., Lau, A., and Kennedy, C. (2018). What matters most for weighting online opt-in samples. Available online at: https://coilink.org/20.500.12592/1zfbv0 (Accessed May 20, 2025).

Ornstein, J. T. (2023). "Getting the most out of surveys: Multilevel regression and poststratification," in *Causality in Policy Studies: a Pluralist Toolbox* (Cham: Springer International Publishing), 99–122. doi: 10.1007/978-3-031-12982-7_5

Pasek, J. (2018). ANESRAKE: ANES raking implementation. R package version 0.80.

Prosser, C., and Mellon, J. (2018). The twilight of the polls? A review of trends in polling accuracy and the causes of polling misses. *Govern. Opposite.* 53, 757–790. doi: 10.1017/gov.2018.7

Rafei, A., Elliott, M. R., and Flannagan, C. A. (2022). Robust and efficient Bayesian inference for non-probability samples. *arXiv preprint arXiv:2203.14355*.

Rafei, A., Flannagan, C. A., and Elliott, M. R. (2020). Big data for finite population inference: applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees. *J. Surv. Statist. Methodol.* 8, 148–180. doi: 10.1093/jssam/smz060

Rivers, D. (2007). "Sampling for web surveys," in *Joint Statistical Meetings, volume 4* (Alexandria, VA: American Statistical Association).

Rocheva, A., Varshaver, E., and Ivanova, N. (2022). "Targeting on social networking sites as sampling strategy for online migrant surveys: the challenge of biases and search for possible solutions," in *Migration Research in a Digitized World*, 35. doi: 10.1007/978-3-031-01319-5_3

Rueda, M., Ferri-García, R., and Castro, L. (2020). The r package nonprobest for estimation in non-probability surveys. *R J.* 12:405. doi: 10.32614/RJ-2020-015

Schirripa Spagnolo, F., Bertarelli, G., Summa, D., Scannapieco, M., Pratesi, M., Marchetti, S., et al. (2025). Inference for big data assisted by small area methods: an application on sustainable development goals sensitivity of enterprises in Italy. *J. R. Stat. Soc. Series A* 188, 27–45. doi: 10.1093/jrsssa/qnae115

Schneider, D., and Harknett, K. (2022). What's to like? Facebook as a tool for survey data collection. *Sociol. Methods Res.* 51, 108–140. doi: 10.1177/0049124119882 477

Schonlau, M., and Couper, M. P. (2017). Options for conducting web surveys. Stat. Sci. 32, 279–292. doi: 10.1214/16-STS597

Selcuki, C. (2023). Why Turkish pollsters didn't foresee Erdogan's win. Available online at: https://foreignpolicy.com/2023/06/07/turkey-elections-polls-erdogan-kilicdaroglu/ (Accessed May 21, 2025).

Shirani-Mehr, H., Rothschild, D., Goel, S., and Gelman, A. (2018). Disentangling bias and variance in election polls. J. Am. Stat. Assoc. 113, 607–614. doi: 10.1080/01621459.2018.1448823

Si, Y. (2020). On the use of auxiliary variables in multilevel regression and poststratification. arXiv preprint arXiv:2011.00360.

Smith, T. (1976). The foundations of survey sampling: a review. J. R. Stat. Soc. 139, 183–195. doi: 10.2307/2345174

Sohlberg, J., Gilljam, M., and Martinsson, J. (2017). Determinants of polling accuracy: the effect of opt-in internet surveys. J. Elect. Public Opin. Part. 27, 433–447. doi: 10.1080/17457289.2017.1300588

Stephan, F. F. (1942). An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Ann. Mathem. Statist.* 13, 166–178. doi: 10.1214/aoms/1177731604

Sturgis, P., Kuha, J., Baker, N., Callegaro, M., Fisher, S., Green, J., et al. (2018). An assessment of the causes of the errors in the 2015 UK general election opinion polls. *J. R. Stat. Soc.* 181, 757–781. doi: 10.1111/rssa.12329

Tam, S.-M., and Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian bureau of statistics. *Int. Stat. Rev.* 83, 436–448. doi: 10.1111/insr. 12105

Tan, Z. (2007). Comment: understanding or, ps and dr. Stat. Sci. 22, 560–568. doi: 10.1214/07-STS227A

Tutz, G. (2023). Probability and non-probability samples: Improving regression modeling by using data from different sources. *Inf. Sci.* 621, 424–436. doi: 10.1016/j.ins.2022.11.032

Unangst, J., Amaya, A. E., Sanders, H. L., Howard, J., Ferrell, A., Karon, S., et al. (2020). A process for decomposing total survey error in probability and nonprobability surveys: a case study comparing health statistics in us internet panels. *J. Surv. Stat. Methodol.* 8, 62–88. doi: 10.1093/jssam/smz040

Valentino, N. A., Zhirkov, K., Hillygus, D. S., and Guay, B. (2020). The consequences of personality biases in online panels for measuring public opinion. *Public Opin. Q.* 84, 446–468. doi: 10.1093/poq/nfaa026

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. J. Surv. Stat. Methodol. 8, 231-263. doi: 10.1093/jssam/smz003

Valliant, R., Dever, J. A., and Kreuter, F. (2013). Practical Tools for Designing and Weighting Survey Samples, volume 1. Cham: Springer. doi: 10.1007/978-1-4614-6449-5

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *Int. J. Forecast.* 31, 980–991. doi: 10.1016/j.ijforecast.2014.06.001

West, B. T., and Andridge, R. R. (2023). Evaluating pre-election polling estimates using a new measure of non-ignorable selection bias. *Public Opin. Q.* 87, 575–601. doi: 10.1093/poq/nfad018

Wu, C. (2022). Statistical inference with non-probability survey samples. Surv. Methodol. 48, 283-311.

Zagheni, E., Garimella, V. R. K., Weber, I., and State, B. (2014). "Inferring international and internal migration patterns from twitter data," in *Proceedings of the 23rd International Conference on World Wide Web*, 439–444. doi: 10.1145/2567948.2576930

Zagheni, E., and Weber, I. (2015). Demographic research with non-representative internet data. Int. J. Manpow. 36, 13–25. doi: 10.1108/IJM-12-2014-0261

Zagheni, E., Weber, I., and Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Popul. Dev. Rev.*721–734. doi: 10.1111/padr.12102

Zarouali, B., Dobber, T., De Pauw, G., and de Vreese, C. (2022). Using a personalityprofiling algorithm to investigate political microtargeting: assessing the persuasion effects of personality-tailored ads on social media. *Communic. Res.* 49, 1066–1091. doi: 10.1177/0093650220961965

Zou, B., Zou, F., Shuster, J. J., Tighe, P. J., Koch, G. G., and Zhou, H. (2016). On variance estimate for covariate adjustment by propensity score analysis. *Stat. Med.* 35, 3537–3548. doi: 10.1002/sim.6943